Similarity Min-Max: Zero-Shot Day-Night Domain Adaptation

Rundong Luo[1,2,3] Wenjing Wang[1] Wenhan Yang[4] Jiaying Liu[1*]
[1] Wangxuan Institute of Computer Technology, Peking University
[2]School of EECS, Peking University [3]School of CS, Peking University [4] Peng Cheng Laboratory

Abstract

Low-light conditions not only hamper human visual experience but also degrade the model's performance on downstream vision tasks. While existing works make remarkable progress on day-night domain adaptation, they rely heavily on domain knowledge derived from the task-specific nighttime dataset. This paper challenges a more complicated scenario with border applicability, i.e., ***zero-shot*** *day-night domain adaptation, which eliminates reliance on any nighttime data. Unlike prior zero-shot adaptation approaches emphasizing either image-level translation or model-level adaptation, we propose a similarity min-max paradigm that considers them under a unified framework. On the image level, we darken images towards minimum feature similarity to enlarge the domain gap. Then on the model level, we maximize the feature similarity between the darkened images and their normal-light counterparts for better model adaptation. To the best of our knowledge, this work represents the pioneering effort in jointly optimizing both levels, resulting in a significant improvement of model generalizability. Extensive experiments demonstrate our method's effectiveness and broad applicability on various nighttime vision tasks, including classification, semantic segmentation, visual place recognition, and video action recognition. Our project page is available at* https://red-fairy.github.io/ZeroShotDayNightDA-Webpage/.

1. Introduction

Deep neural networks are sensitive to insufficient illumination, and such deficiency has posed significant threats to safety-critical computer vision applications. Intuitively, insufficient illumination can be handled by low-light enhancement methods [23, 30, 34, 56, 60, 63], which aim at restoring low-light images to normal-light. However, enhancement models do not necessarily benefit downstream high-level vision tasks as they are optimized for human visual perception and neglect the need for machine vision.

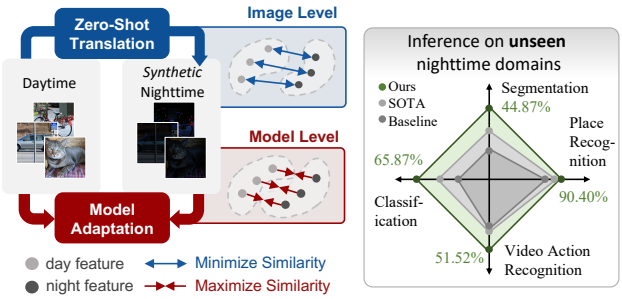


Figure 1. Left: Illustration of our similarity min-max framework for zero-shot day-night domain adaptation. Right: Our framework achieves state-of-the-art results on multiple downstream high-level vision tasks without seeing real nighttime images during training.

Much existing literature has focused on improving machine vision performance at night through domain adaptation. By aligning the distribution statistics between the nighttime and daytime datasets through image translation [2, 12, 45], self-supervised learning [52, 53], or multi-stage algorithms [10, 46, 47], these methods have greatly improved models' performance in nighttime environments. The primary assumption of domain adaptation is that the target domain data is readily available. Nevertheless, obtaining data from the task-specific target domain may be challenging in extreme practical application scenerios such as deep-space exploration and deep-sea analysis.

To reduce the requirement on target domain data, **zero-shot** domain adaptation has emerged as a promising research direction, where adaptation is performed without accessing the target domain. Regarding day-night domain adaptation, the primary challenge is learning illumination-robust representations generalizable to both day and night modalities. To accomplish this goal under zero-shot constraints, Lengyel *et al.* [29] proposed a color invariant convolution for handling illumination changes. Cui *et al.* [8] designed a Reverse ISP pipeline and generated synthetic nighttime images with pseudo labels. However, image-level methods simply consider synthetic nighttime as pseudo-

*Corresponding author.

labeled data and overlook model-level feature extraction; model-level methods focus on adjusting model architecture but neglect image-level nighttime characteristics. Neither is effective enough capture the illumination-robust representations that could bridge the complex day-night domain gap.

From this point of view, we devise a similarity min-max framework that involves two levels, as illustrated in Figure 1. On the image level, we generate a synthetic nighttime domain that shares minimum feature similarity with the daytime domain to enlarge the domain gap. On the model level, we learn illumination-robust representations by maximizing the feature similarity of images from the two domains for better model adaptation.

Intuitive as it seems, solving this bi-level optimization problem is **untrivial**. Directly solving it may yield unsatisfactory results, *e.g.*, meaningless images filled with zero values or identical features given all inputs. Therefore, we develop a stable training pipeline that can be considered a sequential operation on both the image and the model. Regarding the image, we propose an exposure-guided module to perform reliable and controllable nighttime image synthesis. Regarding the model, we align the representation of images from day and night domains through multi-task contrastive learning. Finally, our model achieves day-night adaptation without seeing real nighttime images.

Our framework can serve as a plug-and-play remedy to existing daytime models. To verify its effectiveness, we conduct extensive experiments on multiple high-level nighttime vision tasks, including classification, semantic segmentation, visual place recognition, and video action recognition. Results on various benchmarks demonstrate our superiority over the state-of-the-art.

Our contributions are summarized as follows:

- We propose a similarity min-max framework for zero-shot day-night domain adaptation. Feature similarity between the original and darkened images is minimized by image-level translation and maximized by model-level adaptation. In this way, model's performance in nighttime is improved without accessing real nighttime images.
- We develop a stable training pipeline to solve this bi-level optimization problem. On the image level, we propose an exposure-guided module to perform reliable and controllable nighttime image synthesis. On the model level, we align the representation of images from day and night domains through multi-task contrastive learning.
- Our framework universally applies to various nighttime high-level vision tasks. Experiments on classification, semantic segmentation, visual place recognition, and video action recognition demonstrate the superiority of our method.

2. Related Works

Low-Light Enhancement. A straightforward approach to improve the model's performance in low light is brightening the test low-light images. Early non-learning practices exploit image processing tools such as histogram equalization [40] or image formation theories such as Retinex Theory [44]. Recent literature mainly takes advantage of the advance in deep learning. Trained on paired day-night data, some methods [33, 55, 56] simulate the image decomposition process of Retinex Theory. Others introduce adversarial learning [23] to support unpaired training. Zero-DCE [16, 30] designs a curve-based low-light enhancement model and trains in a zero-reference way. Advanced techniques, including frequency decomposition [24], feature pyramids [60, 63], and flow models [54] are also adopted in recent papers.

Day-Night Domain Adaptation. Nighttime high-level vision has attracted increasing attention in recent years. Apart from pre-processing with enhancement models, day-night domain adaptation is also a viable solution. YOLO-in-the-dark [47] introduces the glue layer to mitigate the day-night domain gap. MAET [8] exploits image signal processing (ISP) for nighttime image generation and uses both synthetic and real nighttime images for training. HLA-face [52] proposes a joint high-low adaptation framework driven by self-supervised learning. Others [2, 37, 45, 46, 57] employ Generative Adversarial Network (GAN) to transfer labeled daytime data to nighttime.

Zero-Shot Day-Night Domain Adaptation. Beyond Conventional adaptation, **zero-shot** approaches consider an even stricter condition where real nighttime images are inaccessible. For general tasks, existing methods either draw supports from extra task-irrelevant source and target domain data pairs [39, 51] or require underlying probability distribution of the target domain [20], which are inapplicable to our settings. For the day-night task, Lengyel *et al.* propose the Color Invariant Convolution (CIConv) [29] to capture illumination-robust features. MAET [8] can be viewed as zero-shot when real nighttime images are discarded during finetuning. Besides, domain generalization methods [1, 6, 19, 26, 31, 38, 62] also apply to our settings since they do not know target domains, but they are too general to handle the complex day-night domain gap.

Despite these advances, low-light enhancement concentrates on human vision and disregards downstream nighttime vision tasks. Conventional adaptation methods require task-specific nighttime datasets, which creates extra burdens on data collection and limits their generalizability to multiple tasks. Prior zero-shot adaptation methods fail to consider image-level and model-level jointly. In this paper, we propose a novel similarity min-max framework that could outperform existing methods by a large margin.

3. Similarity Min-Max Optimization

This section introduces our approach for zero-shot day-night domain adaptation. We first explain our motivation, then introduce the overall framework and detailed designs.

3.1. Motivation

Existing methods, generally categorized into Operator-based and Darkening-based as shown in Figure 2, come across troubles in the day-night domain adaptation problem. Operator-based methods [29] rely on the manually designed operators *at the model level* to handle illumination variations, which are not adaptive to real complex scenarios. Darkening-based methods transfer labeled daytime data to nighttime by ISP [8] or GAN [2, 28, 45, 46] only *at the image level*. However, the former is sensor-dependent and cannot generalize across devices and datasets, while the latter requires data from the task-specific nighttime domain and thus fails to generalize to our zero-shot setting.

Intrinsically, the most critical issue of existing methods is their ignorance of the mutual effect between **pixels** and **features**. In our work, we make the first systematic investigation on this issue and propose a similarity min-max framework that thoroughly exploits the information from two sides. In detail, *at the pixel* (*image*) *level*, we minimize the feature similarity between original and darkened images by day-to-night translation. While *at the feature* (*model*) *level*, we maximize the feature similarity by representation alignment. This joint optimization leads to representations more robust to illumination changes.

We formulate our framework as follows. Denote the feature extractor of the downstream model as $F(\cdot)$. Being robust to illumination requires the extracted feature of a daytime image I and its nighttime version $D(I)$ to be similar, where $D(\cdot)$ represents a darkening process. The limitation of existing darkening-based methods is that their D does not consider the co-effect of F. So we introduce additional constraints on D: we require D to minimize the similarity between the day feature $F(I)$ and the night feature $F(D(I))$. This way, we guide the darkening process with high-level vision, forming a unified framework of D and F. At this point, we can integrate D and F as a min-max optimization problem:

$$\max_{\theta_F} \min_{\theta_D} \quad \mathrm{Sim}(F(I), F(D(I))), \tag{1}$$

where θ_D and θ_F denote the parameters in D and F, and $\mathrm{Sim}(\cdot,\cdot)$ measures the similarity between features.

However, trivial solutions exist in Eq. (1), such as D generating entirely black images and F extracting identical features for all inputs. We add regularizations to D and F accordingly to address this problem:

$$\max_{\theta_F} \min_{\theta_D} \mathrm{Sim}(F(I), F(D(I))) + \mathcal{R}_D(\theta_D) - \mathcal{R}_F(\theta_F), \tag{2}$$

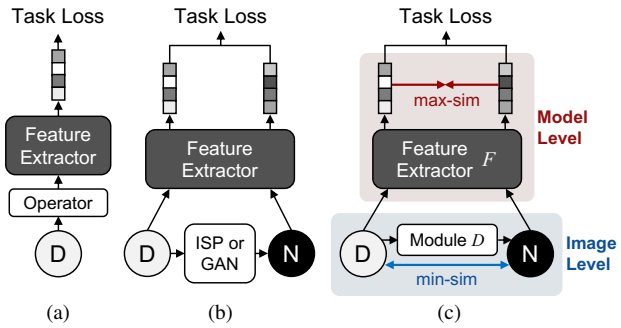


Figure 2. Comparison between different learning paradigms. D and N denote the daytime and nighttime domains, respectively. (a) Operator-based. (b) Darkening-based. (c) Our method.

where $\mathcal{R}_D$ and $\mathcal{R}_F$ are intended to prevent model collapse.

How to design $\mathcal{R}_D$ and $\mathcal{R}_F$ properly is the key to solving Eq. (2). The following will introduce how we design $\mathcal{R}_D$ and $\mathcal{R}_F$ and build up the whole learning framework.

3.2. Image-Level Similarity Minimization

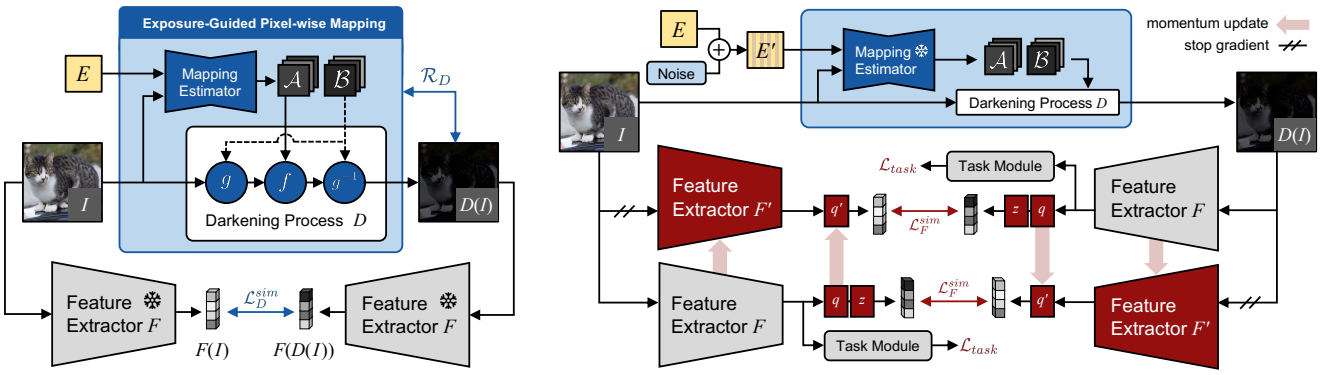
This section describes our design for the darkening module D. We want D to satisfy three properties:

- **Stability**. First and foremost, we need to prevent the similarity min-max optimization from collapsing, *i.e.*, applying proper $\mathcal{R}_D$ in Eq. (2).
- **Generalization**. D should represent a generalized darkening process so the downstream model can learn useful knowledge from $D(I)$ to handle unseen nighttime scenes.
- **Flexibility**. We additionally expect flexible control over the degree of darkening, which could enable us to create diverse inputs beneficial for optimizing F.

We design an exposure-guided pixel-wise mapping algorithm to satisfy the above properties. Unlike widely-used image-to-image darkening approaches [2, 28, 46] that rely heavily on real nighttime images, pixel-wise mapping adjusts images using a pre-selected function with learnable parameters. We empirically found that, by setting proper constraints on the mapping function, we can naturally avoid obtaining trivial solutions in the similarity min-max optimization (*stability*) and guarantee D follows a typical low-light process (*generalization*). Finally, we add an exposure guidance mechanism for better *flexibility*. The detailed design will be illustrated as follows.

Darkening Process. We first define a general function for tone mapping. Given an image $I \in [0,1]^{C\cdot H\cdot W}$, we use a non-linear mapping $f\colon [0,1] \to [0,1]$ and a pixel-wise adjustment map $\mathcal{A} \in [0,1]^{C\cdot H\cdot W}$ to process the image:

$$D^0(I) = f(I, \mathcal{A}). \tag{3}$$



(a) Image-Level Translation

(b) Model-Level Adaptation

Figure 3. Our proposed similarity min-max framework for zero-shot day-night domain adaptation. (a) We first train a darkening module D with a fixed feature extractor to generate *synthesized* nighttime images that share minimum similarity with their daytime counterparts. (b) After obtaining D, we freeze its weights and maximize the day-night feature similarity to adapt the model to nighttime.

Typically, f should be monotonically increasing to preserve contrast and satisfy $f(1, \alpha) = 1$ for all α to avoid information loss (*e.g.*, gamma correction). However, the latter constraint $f(1, \alpha) = 1$ no longer holds for darkening. Therefore, we propose an auxiliary pixel-wise adjustment using a monotonic increasing function g: $[0, 1] \rightarrow [0, 1]$ parameterized by another adjustment map $\mathcal{B} \in [0, 1]^{C \cdot H \cdot W}$. Note that g only serves as a complement and should be simple to avoid taking over the role of f. The overall darkening process is formulated as:

$$D(I) = g^{-1}(f(g(I, \mathcal{B}), \mathcal{A}), \mathcal{B}). \tag{4}$$

Both $\mathcal{A}$ and $\mathcal{B}$ are estimated by a mapping estimator conditioned on the input image I.

To guarantee D represents a darkening process (*i.e.*, $D(I) < I$), f should additionally satisfy convexity. Specifically, we let f be the iterative quadratic curve [16]: $f(x) = h^{(8)}(x)$, $h(x, \alpha) = \alpha x^2 + (1 - \alpha)x$, and g be the dividing operation: $g(x, \beta) = x/\beta$ in our implementation. Other kinds of curve forms are also considered and tested. Still, we empirically found that quadratic curves could bring slightly better results (results in Sec. 4.2).

Besides, to enable flexible control over the exposure level, we feed an exposure map E to the mapping estimator with I, yielding the corresponding darkened image $D(I, E)$. During training, the darkening module is encouraged to align the pixel value of E and $D(I, E)$. We use $D(I)$ and $D(I, E)$ interchangeably for simplicity.

Similarity Minimization. The training objective of module D involves two parts: similarity minimization and regularization. For the former, we directly reduce the distance between features:

$$\mathcal{L}_D^{sim} = \frac{\langle F(I), F(D(I)) \rangle}{||F(I)||_2 \cdot ||F(D(I))||_2}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between two vectors.

The regularization term consists of four losses. Besides a color consistency loss $\mathcal{L}_{col}$ [16] that corrects color deviations, three additional losses are proposed to regularize D:

Firstly, conditional exposure control is adopted to align the exposure map with the corresponding generated image:

$$\mathcal{L}_{c-exp} = \sum_{1 \le i \le H, 1 \le j \le W} |\hat{D}_{i,j}(I, E) - E_{i,j}|, \tag{6}$$

where $\hat{D}(I, E)$ is the channel-wise average of $D(I, E)$. During training, each exposure map E has identical entries uniformly sampled between $[0, 0.5]$.

Then we add constraints on $\mathcal{A}$. Intuitively, $\mathcal{A}$ represents the degree of illumination reduction. Illumination usually varies slowly across a scene but encounters rapid variations from object to object. Following this property, we apply a loose total variance loss:

$$\mathcal{L}_{ltv}(\mathcal{A}) = \sum_{c \in \{R,G,B\}} (h(|\nabla_x \mathcal{A}^c|)^2 + h(|\nabla_y \mathcal{A}^c|)^2), \tag{7}$$

$$h(x) = \max(\alpha - |x - \alpha|, 0), \tag{8}$$

where ∇_x, ∇_y are gradient operations along the horizontal and vertical axis, respectively, and α is a hyperparameter. Compared with the original total variance loss where h is the identity function, our loose version allows the network to predict values of greater difference for adjacent pixels, which is common on objects' boundaries.

Finally, we adopt $\mathcal{L}_{flex}(\mathcal{B}) = 1 - \mathcal{B}$ to avoid model fitting to the exposure solely by g.

The overall training objective for D is:

$$\mathcal{L}_D = \lambda_D^{sim}\mathcal{L}_D^{sim} + \mathcal{R}_D, \tag{9}$$

$$\mathcal{R}_D = \lambda_{c-exp}\mathcal{L}_{c-exp} + \lambda_{col}\mathcal{L}_{col} + \lambda_{ltv}\mathcal{L}_{ltv} + \lambda_{flex}\mathcal{L}_{flex}. \tag{10}$$

3.3. Model-Level Similarity Maximization

The darkening module D grants us access to a synthetic nighttime domain. In this section, we exploit D to learn illumination-robust representations.

Contrastive learning [5, 17] is a self-supervised learning paradigm that contrasts positive and negative image pairs. However, images of the same class in classification or adjacent scenes in segmentation will form false negative pairs, thus hurting the model's performance. To alleviate these burdens, BYOL [15] proposes a non-negative variant that only aligns the feature between positive image pairs $\{v, v^+\}$:

$$\mathcal{L}_{\text{BYOL}}(v, v^+) = 2 - \frac{2 \cdot \langle z(q(F(v))), q'(F'(v^+))\rangle}{||z(q(F(v)))||_2 \cdot ||q'(F'(v^+))||_2}, \tag{11}$$

where q, q' are projection heads, and z is the prediction head. Both of them are MLPs with a single hidden layer. Note that F' and q' share the same architecture and weight initialization with F and q but receive no gradient and are updated by exponential moving average (EMA).

Similarity Maximization. Motivated by BYOL, we maximize the feature similarity between synthetic nighttime and daytime domains by non-negative contrastive learning. Given a daytime image I and an exposure map E, we formulate the training objective as follows:

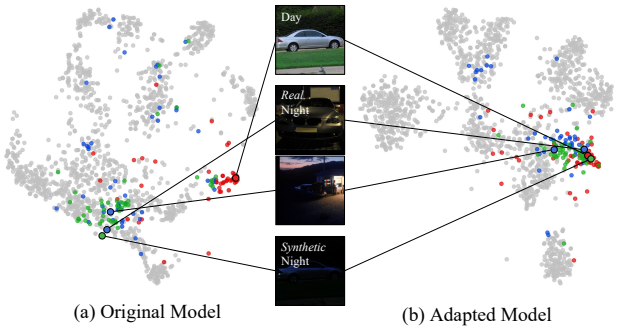
$$\mathcal{L}_F^{sim} = \mathcal{L}_{\text{BYOL}}(I, D(I, E)) + \mathcal{L}_{\text{BYOL}}(D(I, E), I). \tag{12}$$

Note that the measure of feature similarity is different between Eq. (5) and Eq. (12). Directly applying Eq. (5) to train F brings poorer results due to potential feature degeneration. In comparison, the asymmetric projection head and stop gradient policies prevent the feature extractor F from collapsing, *i.e.*, working as the regularization $\mathcal{R}_F$ in Eq. (2) together with the task loss (introduced below).

Moreover, different from E in Eq. (6), we use a compound exposure map E' instead. E' is first initialized with identical entries uniformly sampled between $[0, 0.2]$ for simulating nighttime illumination. This range is the same for all downstream tasks, which does not introduce task-relevant prior. Then, we add pixel-wise noise z_1 and patch-wise noise z_2 to E to simulate exposure discrepancy. Overall, E' can be represented as:

$$E' = \mathcal{U}(0, 0.2) + z_1 + z_2. \tag{13}$$

See the supplementary for details on noise injection.



(a) Original Model (b) Adapted Model

Figure 4. t-SNE [50] visualization of images' feature extracted by the original daytime model and our adapted model on CODaN [29]. Red, green, and blue dots represent the feature of daytime, *synthesized* nighttime, and *real* nighttime images, respectively. We only color the instances from the "Car" category for better visual quality. Additional visualization results are shown in the supplementary.

Besides $\mathcal{L}_F^{sim}$, we add task-specific supervision $\mathcal{L}_{task}$ on both the original daytime and synthetic nighttime domain. The final training objective for F is:

$$\mathcal{L}_F = \lambda_F^{sim}\mathcal{L}_F^{sim} + \lambda_{task}\mathcal{L}_{task}. \tag{14}$$

3.4. Overall Training Pipeline

Having introduced the image-level similarity minimization (Section 3.2) and model-level similarity maximization (Section 3.3), this section discusses the overall pipeline, as shown in Figure 3.

An intuitive idea is training D and F alternately like GAN [13, 64]. Nevertheless, balancing D and F increases the difficulty of parameter tuning and makes the optimization process unstable. We adopt a simple but effective two-step strategy to solve this problem: we first train D and keep F frozen, then train F and keep D frozen. Compared with the alternate strategy, our step-wise approach improves the performance on nighttime image classification (elaborate in Section 4.2) from 63.84% to 65.87%.

We could also explain the merits of our min-max framework from the perspective of adversarial training [14, 35]. Module D first produces the worst-case examples regarding feature similarity. Then, our model could learn the illumination-robust features by learning on these cases through similarity maximization. This technical commonality further justifies our motivation to build the similarity min-max framework.

Across all downstream tasks, the feature extractor and task module are initialized by daytime pre-trained models. We first freeze the feature extractor and train the darkening module (image-level translation). Then, we keep the darkening module fixed and train the feature extractor and task module jointly (model-level adaptation).

3.5. Empirical Justifications on Darkening Module

Simulating nighttime conditions without accessing real nighttime images is the key to our framework. Particularly, nighttime conditions bring semantic changes in addition to illumination changes, *e.g.*, the dark environment with artificial lights on the second real nighttime image in Figure 4. However, an accurate simulation is extremely difficult since our prior knowledge is limited to "low illumination". Fortunately, unlike typical day-to-night image synthesis processes [41] which target the human visual experience, ours only care about the distribution of darkened images in the feature space. Leaving aside visual quality, we are pleased to find that the feature distribution of our synthesized nighttime domain is similar to that of the real nighttime domain as visualized in Figure 4(a). This observation demonstrates that our darkening process can characterize the night domain from the model-level perspective.

Thanks to this property, the feature discrepancy between daytime and *real* nighttime domain is significantly reduced after model-level adaptation (red and blue dots in Figure 4). This discovery is consistent with the Maximum Mean Discrepancy (MMD) between the feature distribution of day and night modalities, which is 0.020 and 0.014 for the original and adapted models, respectively. We provide implementation details and additional empirical analysis using saliency maps in the supplementary.

4. Experiments

This section provides the implementation details, benchmarking results, and ablation analysis of our method.

4.1. Implementation Details

Our framework widely applies to various nighttime vision tasks. In the following, we evaluate our method with four representative tasks: image classification, semantic segmentation, visual place recognition, and video action recognition. Only daytime data are accessible for training and validation, while nighttime data are only used during evaluation. We benchmark our method with three categories of methods that require no dataset-specific target domain data: low-light enhancement, zero-shot day-night domain adaptation, and domain generalization. For low-light enhancement, enhancement models are trained on their original datasets. Then we adopt them as a pre-processing step to assist the daytime baseline. The results of our method is the average of three independent trails. Additional details are provided in the supplementary.

4.2. Nighttime Image Classification

We first consider one of the most fundamental vision tasks: image classification. CODaN [29] is a 10-class dataset containing a training set of 10000 daytime images and a test set with 2500 daytime and nighttime images, respectively. We validate models on the daytime test set and evaluate them on the nighttime test set. The backbone is ResNet-18 [18].

Table 1. Top-1 classification accuracy on the CODaN nighttime test set [29]. † denotes our re-implementation with both the original and synthesized image fed into the task module.

Method	Top-1 (%)
ResNet-18 [18]	53.32
Low-Light Enhancement	
EnlightenGAN [23]	56.68
LEDNet [63]	57.40
Zero-DCE++ [30]	57.96
RUAS [33]	58.36
SCI [34]	58.68
URetinexNet [56]	58.72
Domain Generalization	
MixStyle [62]	53.12
IRM [1]	54.52
AdaBN [31]	54.25
Zero-Shot Day-Night Domain Adaptation	
MAET† [8]	56.48
CIConv [29]	60.32
Ours	**65.87**

Benchmarking results are shown in Table 1. Enhancement methods restore input low-light images from the human visual perspective while keeping the model untouched, resulting in limited performance gains. Domain generalization methods are designed for general tasks and perform poorly in unseen nighttime environments. MAET [8] relies on degrading transformation with sensor-specific parameters, which suffers from poor generalizability. CIConv [29] adopts learnable color invariant edge detectors, which are not robust to the complex illumination variation in real scenarios. In contrast, our method outperforms state-of-the-art methods by a large margin (60.32% v.s. 65.87%), demonstrating our unified framework could obtain features more robust to illumination shifts.

Ablation Studies. We conduct ablation studies to justify our framework design in Table 2. Firstly, we study how to design the darkening module D given $\mathcal{L}_D^{sim}$. The model-level adaptation stage (Section 3.3) remains the same for fair comparisons. Firstly, we replace our darkening module with heuristic image adjustment approaches, such as brightness adjustment (`Brightness` in `PIL`[1]) and gamma correction ($D(I) = I^\gamma$). We implement these two approaches using a fixed darkening hyperparameter chosen after multi-

[1]https://pillow.readthedocs.io/en/stable/reference/ImageEnhance.html

Table 2. Ablation studies for module D and similarity losses. We report the Top-1 accuracy on the CODaN [29] nighttime test set.

Category	Method	Top-1 (%)
Baseline	Vanilla ResNet-18	53.32
Module D Heuristic	Brightness adjustment	57.96
	Gamma correction	63.96
Module D Learnable	Reciprocal curve	62.60
	Gamma curve	64.16
Similarity Loss	w/o $\mathcal{L}_D^{sim}$ and $\mathcal{L}_F^{sim}$	64.08
	w/o $\mathcal{L}_D^{sim}$	64.56
	w/o $\mathcal{L}_F^{sim}$	64.88
Full version	-	**65.87**

ple trials and report the best score.

Next, we test other possible curve forms for f. Both gamma curve ($f(x, \alpha) = x^{\frac{1}{\alpha}}, \alpha \in (0, 1]$) and reciprocal curve ($f(x, \alpha) = \frac{(1-\alpha)\cdot x}{1-\alpha \cdot x}, \alpha \in [0, 1)$) bring slightly worse results than the iterative quadratic curve. Please refer to the supplementary for implementation details of these ablations and additional results on the segmentation task.

Finally, we test our framework's performance when one or both of the similarity loss is absent. We find that either similarity loss alone can boost the model's nighttime performance while combining them achieves the best result.

4.3. Nighttime Semantic Segmentation

Next, we explore a more challenging nighttime vision task: semantic segmentation. We adopt RefineNet [32] with ResNet-101 backbone as the baseline. The daytime training dataset is Cityscapes [7], containing 2975 images for training and 500 images for validation, all with dense annotations. The nighttime testing datasets are Nighttime Driving [9] and Dark-Zurich [46]. These two datasets contain 50 coarsely annotated and 151 densely annotated nighttime street view images.

We benchmark our method in Table 3. Low-light enhancement methods yield worse results than the baseline because they perform poorly on street scenes with complex light sources. Domain generalization methods fail to mitigate the huge day-night domain gap, leading to unsatisfactory results. Note that RobustNet [6] adopts DeepLab-v3 [4] architecture, which is superior to RefineNet [32] adopted in our implementation. Among zero-shot adaptation methods, MAET [8] injects too much noise into images, leading to severe performance degradation. CIConv yields better results, but the improvement is limited. In comparison, our approach improves the mIoU scores to 44.9% on Nighttime Driving and 40.2% on Dark-Zurich.

Figure 5 shows qualitative segmentation results on two nighttime datasets. Low-light enhancement methods perform poorly on nighttime street scenes. Our method better extracts information hidden by darkness and thus generates more accurate semantic maps.

Table 3. Semantic segmentation results on Nighttime Driving [9] and Dark-Zurich [46], reported as percentage mIoU scores.

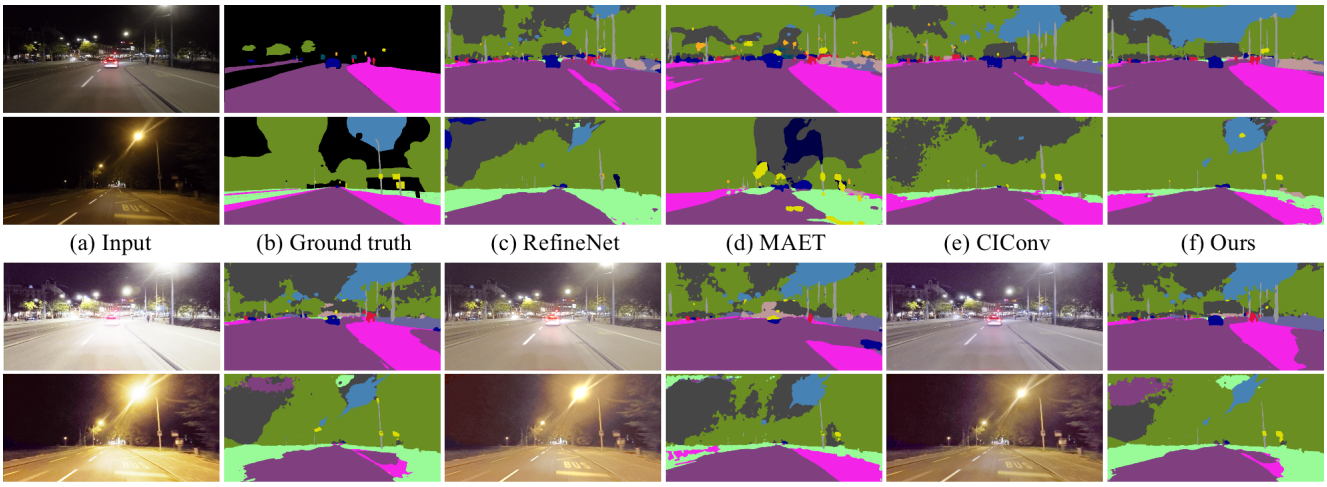
Method	Nighttime Driving	Dark-Zurich
RefineNet [32]	34.3	30.6
Low-Light Enhancement		
EnlightenGAN [23]	25.2	24.9
Zero-DCE++ [30]	32.7	28.3
RUAS [33]	25.1	23.4
SCI [34]	28.6	25.7
URetinexNet [56]	28.1	24.0
LEDNet [63]	27.6	26.6
Domain Generalization		
AdaBN [31]	37.2	31.1
RobustNet [6]	33.0	34.5
SAN-SAW [38]	28.1	16.0
Zero-Shot Day-Night Domain Adaptation		
MAET [8]	28.1	26.4
CIConv [29]	41.2	34.5
Ours	**44.9**	**40.2**

Table 4. Visual place recognition results on Tokyo 24/7 [49].

Method	mAP (%)
Zero-Shot Day-Night Domain Adaptation	
EdgeMAC [42]	75.9
U-Net jointly [21]	79.8
GeM [43]	85.0
CIConv-GeM [29]	88.3
Ours-GeM	**90.4**
Day-Night Domain Adaptation (night images are available for training)	
U-Net jointly [21]	86.5
EdgeMAC + CLAHE [21]	90.5
EdgeMAC + U-Net jointly [21]	90.0

4.4. Visual Place Recognition at Night

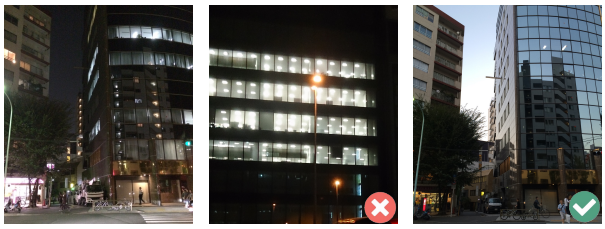
Then we explore visual place recognition, which aims to retrieve images that illustrate the same scene of a query image from an image pool. Unlike classification and segmentation, place recognition methods are not end-to-end during inference. We extend our method based on GeM [43] with ResNet-101 backbone. In GeM, the network receives a tuple of images $\{p, q, n_1, \cdots, n_k\}$ as input, in which the query q only matches p. The network is trained on a contrastive loss, similar to the model-level stage in our



(a) Input (b) Ground truth (c) RefineNet (d) MAET (e) CIConv (f) Ours

(g) SCI (h) URetinexNet (i) Zero-DCE++

Figure 5. Semantic segmentation results. For each group, the first row: Nighttime Driving [9], the second row: Dark-Zurich [46].



(a) Query (b) GeM (c) Ours

Figure 6. Qualitative visual place recognition results. (a) A night query from the Tokyo 24/7 dataset [49]. (b) Image retrieved by GeM [43]. (c) Image retrieved by our method.

framework. We retain the image-level stage and modify the model-level stage in our implementation. We first train the darkening module D as usual. Then, we consider $D(p)$ as an additional matching for p, *i.e.*, an input tuple contains two positive samples (instead of one) and k negative samples. We train our network on the Retrieval-SfM dataset [43] and evaluate it on the Tokyo 24/7 dataset [49], which contains city views in multiple illumination conditions and viewing directions.

Performance is reported as mean Average Precision (mAP) in Table 4. Results of comparison methods are borrowed from [21] and [29]. Our method outperforms all zero-shot methods and is comparable to conventional domain adaptation methods. As shown in Figure 6, the baseline method gets fooled by the night's appearance, while our model finds the correct daytime image.

4.5. Low-Light Video Action Recognition

Although initially designed for images, our method also applies to video tasks. Here we consider an 11-class low-light video action recognition task. Normal light training data consists of 2.6k normal light video clips from HMDB51 [27], UCF101 [48], Kinetics-600 [25], and Moments in Time [36]. We evaluate our model on the official test split of the ARID dataset [58]. The action recognizer is I3D [3] based on 3D-ResNet [11].

Table 5. Video action recognition results on ARID [58].

Method	Top-1 (%)
I3D [3]	47.02
Low-Light Video Enhancement	
StableLLVE [59]	45.08
SMOID [22]	47.27
SGZ [61]	46.42
Domain Generalization & Zero-Shot Day-Night Domain Adaptation	
AdaBN [31]	46.17
Ours	**51.52**

We extend our method to video as follows. When training the darkening module, we input frames extracted from video clips. $\mathcal{A}$ and $\mathcal{B}$ in Eq. (4) is estimated for every individual frame. We calculate $\mathcal{L}_D^{sim}$ between video clips and other losses between frames. When generating low-light videos, frames are separately fed into the curve estimator while sharing the same exposure map E'.

We report the results as Top-1 accuracy. As shown in Table 5, video enhancement methods StableLLVE [59], SMOID [22], and SGZ [61] yield a limited performance gain. Meanwhile, our approach boosts models' performance by 4.38%, demonstrating our superiority on videos.

5. Conclusion

In this paper, we propose a novel approach for zero-shot day-night domain adaptation. Going beyond a simple focus on the image-level translation or model-level adaptation, we observe a complementary relationship between two aspects and build our framework upon the similarity min-max paradigm. Our proposed method can significantly boost the model's performance at nighttime without accessing the nighttime domain. Experiments on multiple datasets demonstrate the superiority and broad applicability of our approach.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China under Contract No.61772043 and a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). This research work is also partially supported by the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv*, 2019. 2, 6

[2] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *IJCNN*, 2019. 1, 2, 3

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 8

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017. 7

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 5

[6] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 2, 7

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7

[8] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *ICCV*, 2021. 1, 2, 3, 6, 7

[9] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, 2018. 7, 8

[10] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. Nightlab: A dual-level architecture with hardness detection for segmentation at night. In *CVPR*, 2022. 1

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 8

[12] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, 2022. 1

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 5

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 2014. 5

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, C. Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 5

[16] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020. 2, 4

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, 2021. 2

[20] Masato Ishii, Takashi Takenouchi, and Masashi Sugiyama. Zero-shot domain adaptation based on attribute information. In *ACML*, 2019. 2

[21] Tomas Jenicek and Ondrej Chum. No fear of the dark: Image retrieval under varying illumination conditions. In *ICCV*, 2019. 7, 8

[22] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, 2019. 8

[23] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 2021. 1, 2, 6, 7

[24] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *ECCV*, 2022. 2

[25] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, 2017. 8

[26] Namyup Kim, Taeyoung Son, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: web-image assisted domain generalization for semantic segmentation. *arXiv*, 2021. 2

[27] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011. 8

[28] Hongjun Lee, Moonsoo Ra, and Whoi-Yul Kim. Nighttime data augmentation using gan for improving blind-spot detection. *IEEE Access*, 2020. 3

[29] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *ICCV*, 2021. 1, 2, 3, 5, 6, 7, 8

[30] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE TPAMI*, 2021. 1, 2, 6, 7

[31] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *ICLRW*, 2017. 2, 6, 7, 8

[32] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 7

[33] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 2, 6, 7

[34] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 1, 6, 7

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv*, 2017. 5

[36] Mathew Monfort, Carl Vondrick, Aude Oliva, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa M. Brown, Quanfu Fan, and Dan Gutfreund. Moments in time dataset: One million videos for event understanding. *IEEE TPAMI*, 2020. 8

[37] Amitangshu Mukherjee, Ameya Joshi, Anuj Sharma, Chinmay Hegde, and Soumik Sarkar. Generative semantic domain adaptation for perception in autonomous driving. *Journal of Big Data Analytics in Transportation*, 2022. 2

[38] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *CVPR*, 2022. 2, 7

[39] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *ECCV*, 2018. 2

[40] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 1987. 2

[41] Abhijith Punnappurath, Abdullah Abuolaim, Abdelrahman Abdelhamed, Alex Levinshtein, and Michael S Brown. Day-to-night image synthesis for training nighttime neural isps. In *CVPR*, 2022. 6

[42] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Deep shape matching. In *ECCV*, 2018. 7

[43] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 2019. 7, 8

[44] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Retinex processing for automatic image enhancement. *Journal of Electronic Imaging*, 2004. 2

[45] Eduardo Romera, Luis M. Bergasa, Kailun Yang, Jose M. Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. *IEEE Intelligent Vehicles Symposium*, 2019. 1, 2, 3

[46] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019. 1, 2, 3, 7, 8

[47] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *ECCV*, 2020. 1, 2

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 8

[49] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 7, 8

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 5

[51] Jinghua Wang and Jianmin Jiang. Conditional coupled generative adversarial networks for zero-shot domain adaptation. In *ICCV*, 2019. 2

[52] Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Unsupervised face detection in the dark. *IEEE TPAMI*, 2022. 1, 2

[53] Wenjing Wang, Zhengbo Xu, Haofeng Huang, and Jiaying Liu. Self-aligned concave curve: Illumination enhancement for unsupervised adaptation. In *ACM MM*, 2022. 1

[54] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI*, 2022. 2

[55] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2017. 2

[56] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 1, 2, 6, 7

[57] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, 2021. 2

[58] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition*, 2021. 8

[59] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 2021. 8

[60] Yuzhi Zhao, Yongzhe Xu, Qiong Yan, Dingdong Yang, Xuehui Wang, and Lai-Man Po. D2hnet: Joint denoising and deblurring with hierarchical network for robust night image restoration. In *ECCV*, 2022. 1, 2

[61] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *WACVW*, 2022. 8

[62] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 2, 6

[63] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 1, 2, 6, 7

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 5